Title

Domain-General and Domain-Specific Patterns of Activity Supporting Metacognition in Human Prefrontal Cortex.

Journal

The Journal of neuroscience : the official journal of the Society for Neuroscience, 38(14)

ISSN

0270-6474

Authors

Morales, Jorge
Lau, Hakwan
Fleming, Stephen M

Publication Date

2018-04-01

DOI

10.1523/jneurosci.2360-17.2018

Peer reviewed

Behavioral/Cognitive

Domain-General and Domain-Specific Patterns of Activity Supporting Metacognition in Human Prefrontal Cortex

Jorge Morales,[1] Hakwan Lau,[2,3,4] and Stephen M. Fleming[5,6]

[1]Department of Philosophy, Columbia University, New York, New York 10027, [2]Department of Psychology, University of California, Los Angeles, California 90095, [3]Brain Research Institute, University of California, Los Angeles, California 90095, [4]Department of Psychology, University of Hong Kong, Hong Kong, [5]Wellcome Centre for Human Neuroimaging, University College London, London, United Kingdom WC1N 3BG, and [6]Max Planck Centre for Computational Psychiatry and Ageing Research, University College London, London, United Kingdom WC1B 5EH

Metacognition is the capacity to evaluate the success of one's own cognitive processes in various domains; for example, memory and perception. It remains controversial whether metacognition relies on a domain-general resource that is applied to different tasks or if self-evaluative processes are domain specific. Here, we investigated this issue directly by examining the neural substrates engaged when metacognitive judgments were made by human participants of both sexes during perceptual and memory tasks matched for stimulus and performance characteristics. By comparing patterns of fMRI activity while subjects evaluated their performance, we revealed both domain-specific and domain-general metacognitive representations. Multivoxel activity patterns in anterior prefrontal cortex predicted levels of confidence in a domain-specific fashion, whereas domain-general signals predicting confidence and accuracy were found in a widespread network in the frontal and posterior midline. The demonstration of domain-specific metacognitive representations suggests the presence of a content-rich mechanism available to introspection and cognitive control.

Key words: **confidence; fMRI; memory; metacognition; MVPA; perception**

Significance Statement

We used human neuroimaging to investigate processes supporting memory and perceptual metacognition. It remains controversial whether metacognition relies on a global resource that is applied to different tasks or if self-evaluative processes are specific to particular tasks. Using multivariate decoding methods, we provide evidence that perceptual- and memory-specific metacognitive representations coexist with generic confidence signals. Our findings reconcile previously conflicting results on the domain specificity/generality of metacognition and lay the groundwork for a mechanistic understanding of metacognitive judgments.

Introduction

Metacognition is the capacity to evaluate the success of one's cognitive processes in various domains; for example, perception or memory (Flavell, 1979; Nelson and Narens, 1990; Metcalfe and Shimamura, 1994; Fleming et al., 2012a). Metacognitive ability can be assessed in the laboratory by quantifying the trial-by-trial correspondence between objective performance and subjective confidence (Galvin et al., 2003; Maniscalco and Lau, 2012; Overgaard and Sandberg, 2012; Fleming and Lau, 2014). Anatomical (Fleming et al., 2010, 2012b; McCurdy et al., 2013; Maniscalco et al., 2017), functional (Fleck et al., 2006; Yokoyama et al., 2010; Baird et al., 2013; Hilgenstock et al., 2014), and neuropsychological (Shimamura and Squire, 1986; Schnyer et al., 2004; Fleming et al., 2014) evidence indicates specific neural substrates (especially in frontolateral, frontomedial, and parietal regions) contribute to metacognition across a range of task domains, including perception and memory. However, the neurocognitive architecture supporting metacognition remains controversial. Does metacognition rely on a common, domain-general resource that is recruited to evaluate performance on a variety of tasks? Or is metacognition supported by domain-specific components?

Current computational perspectives (Pouget et al., 2016; Fleming and Daw, 2017) suggest that both domain-general and domain-

Received Aug. 19, 2017; revised Jan. 8, 2018; accepted Jan. 10, 2018.

Author contributions: J.M. and S.M.F. wrote the first draft of the paper; J.M., H.L., and S.M.F. edited the paper; J.M., H.L., and S.M.F. designed research; J.M. and S.M.F. performed research; J.M., H.L., and S.M.F. contributed unpublished reagents/analytic tools; J.M., H.L., and S.M.F. analyzed data; J.M., H.L., and S.M.F. wrote the paper.

This work was supported by the National Institute of Neurological Disorders and Stroke of the National Institutes of Health (Grant R01NS088628 to H.L. and S.M.F.), the Templeton Foundation (Grant 21569 to H.L.), and the Wellcome Trust (Grant WT096185 to S.M.F.). The Wellcome Centre for Human Neuroimaging is supported by core funding from the Wellcome Trust (Grant 203147/Z/16/Z). We thank Aurelio Cortese for help with the MVPA analyses.

The authors declare no competing financial interests.

Correspondence should be addressed to Jorge Morales, Department of Philosophy, Columbia University, 708 Philosophy Hall, 1150 Amsterdam Ave., New York, NY 10027. E-mail: jorge.morales@columbia.edu.

DOI:10.1523/JNEUROSCI.2360-17.2018

specific representations may be important for guiding behavior. One needs to be able to compare confidence estimates in a "common currency" across a range of arbitrary decision scenarios (de Gardelle and Mamassian, 2014). One solution to this problem is to maintain a global resource with access to arbitrary sensorimotor mappings (Holroyd et al., 2005; Heekeren et al., 2006; Cole et al., 2013). Candidate neural substrates for a domain-general resource are the frontoparietal and cingulo-opercular networks, known to be involved in arbitrary control operations (Cole et al., 2013). In particular, the posterior medial prefrontal cortex (PFC) (encompassing the paracingulate cortex and presupplementary motor area) has been implicated in representing confidence, monitoring conflict and detecting errors across a range of tasks (Gehring et al., 1993; Botvinick et al., 2004; Ridderinkhof et al., 2004; Fleming et al., 2012b). Conversely, if the system only had access to generic confidence signals, then appropriate switching between particular tasks or strategies on the basis of their expected success would be compromised. Functional imaging evidence implicates human anterior PFC in tracking the reliability of specific alternative strategies during decision making (Donoso et al., 2014) and such regions may also support domain-specific representations of confidence.

Current behavioral evidence of a shared resource for metacognition is ambiguous, in part due to the difficulty of distilling metacognitive processes from those supporting primary task performance (Galvin et al., 2003; Maniscalco and Lau, 2012; Fleming and Lau, 2014). Some studies have found that efficient metacognition in one task predicts good metacognition in another (McCurdy et al., 2013; Ais et al., 2016; Ruby et al., 2017; Samaha and Postle, 2017; Faivre et al., 2018), whereas others indicate the independence of metacognitive abilities (Kelemen et al., 2000; Baird et al., 2013; Vo et al., 2014). Recent studies using bias-free measures of metacognition have identified differences in the neural correlates of memory and perceptual metacognition in both healthy subjects (Baird et al., 2013; McCurdy et al., 2013) and neuropsychological patients (Fleming et al., 2014). However, the study of behavioral individual differences provides only indirect evidence of the neural and computational architecture supporting metacognition.

Here, we investigated this ontology directly by examining neural substrates engaged when metacognitive judgments are made during perceptual and memory tasks matched for stimulus and performance characteristics. We used a combination of univariate and multivariate analyses of fMRI data to identify domain-specific and domain-general neural substrates engaged during metacognitive judgments. We also distinguished activations engaged by a metacognitive judgment from neural activity that tracks confidence level. Together, our findings reveal the coexistence of generic and specific confidence representations consistent with a computational hierarchy underpinning effective metacognition.

Materials and Methods

Participants

Thirty healthy subjects (ages 18–33 years, mean 24.97; SD = 4.44; 14 males) with normal or corrected-to-normal vision were monetarily compensated and gave written informed consent to participate in the study at the Center for Neural Science at New York University. The study protocols were approved by the local institutional review board. The number of participants was determined a priori at $n = 30$, which is consistent with recent guidelines on neuroimaging sample sizes (Poldrack et al., 2017). Due to behavioral and in-scanner motion cutoff criteria, six subjects were excluded from further analysis (details below). We present the results of 24 subjects whose data were fully analyzed.

Experimental and task design

The experiment had a $2 \times 2 \times 2$ design: condition (confidence/follow) $\times$ task domain (perception/memory) $\times$ stimulus type (shapes/words). It consisted of six scanner runs, each with eight nine-trial miniblocks (72 trials per run, 432 trials in total). Perceptual and memory two-alternative forced-choice (2-AFC) tasks were presented in separate, interleaved runs (three runs per task; order counterbalanced across subjects). In each run, there were four pairs of miniblocks from the confidence and follow conditions. To avoid stimulus confounds, two different types of stimulus were used throughout the experiment. In each run, two pairs of confidence/follow miniblocks used words and the remaining two pairs used abstract shapes (interleaved and order counterbalanced across runs).

In the perceptual task, subjects were asked to report the brighter of two stimuli on each trial. In the memory task, subjects began each miniblock by learning a set of nine consecutively presented stimuli. A stimulus from this set was then presented on each subsequent trial (in randomized order) alongside a new stimulus. The subjects' task was to identify the studied stimulus. In miniblocks from the confidence condition, subjects had to rate their confidence in their performance in each trial by selecting a number from a scale of 1 to 4. In miniblocks from the follow condition, subjects had to "follow the computer" in each trial by pressing the button corresponding to the highlighted number regardless of their confidence. The highlighted number was yoked to their ratings in the previous confidence miniblock (randomized presentation order) to ensure similar low-level visuomotor characteristics in both conditions for any given pair of miniblocks.

Subjects were reminded at the beginning of each miniblock of the condition, task, and stimulus type that would follow. They used two fingers of their right hand to respond on an MRI-compatible button box: left stimulus (index) and right stimulus (middle). For confidence ratings, they used four fingers: 1 (index), 2 (middle), 3 (ring), and 4 (little). If subjects failed to provide either type of response within the allotted time (see Fig. 1*A* for details), the trial was missed and an exclamation mark was displayed for the remainder of the trial. Failing to press the highlighted number counted as a missed trial.

Before entering the scanner, participants were familiarized with the tasks and the confidence rating scale. After computing independent brightness thresholding for words and abstract shapes, subjects practiced one of each miniblock type (i.e., eight miniblocks). Instructions emphasized that confidence ratings should reflect relative confidence and participants were encouraged to use all ratings. The whole experiment lasted ~1.5 h.

Stimuli

The experiment was programmed in MATLAB (The MathWorks) and stimuli were presented using Psychtoolbox (Brainard, 1997). Abstract, 22- or 28-line shapes were created randomly by specifying an (invisible) grid of 6×6 squares that subtended 4 degrees of visual angle where lines could connect two vertices horizontally, vertically, or diagonally. The first line always stemmed from the central vertex of the invisible grid randomly connecting one of the surrounding eight vertices to ensure shape centrality within the grid. The remaining lines were drawn sequentially, ensuring that all lines were connected. Orientation and originating vertices were selected randomly.

All words were nouns of six to 12 letters with one to four syllables obtained from the Medical Research Council Psycholinguistic Database (Wilson, 1988). In the perceptual task, words had high familiarity, concreteness, and imageability ratings (400–700). In the memory task, words had low ratings (100–400) to increase task difficulty. Each word and each shape was presented once throughout the experiment (across perceptual and memory blocks, including practice trials). All subjects were tested on the same words and shapes (counterbalanced across confidence and follow conditions across subjects). Words and rating scales were presented using DS-Digital font (40 points) to make their visual features similar to the abstract shapes.

To obtain stimulus sets of similar difficulty for shapes and words, we ran a series of pilot studies in which participants rated abstract shapes' distinctiveness and then performed the memory task [15 miniblocks per subject; 171 Amazon Mechanical Turk participants (73 for shapes; 98 for

words) and six subjects in the laboratory who performed a complete version of the experiment]. Based on these results, we expected a mean performance in the memory task of ~71% correct responses when 22- and 28-line distinctive shapes were used in the same block and ~83% correct when long words (6–12 letters) with low concreteness, imageability, and familiarity ratings (100–400) were used. To further increase difficulty, we created pairs of old and new words split between the confidence and follow conditions (counterbalanced across subjects), blocked by similar semantic category (e.g., finance, argumentation, character traits, etc.), such that each new word within a block was freely associated with one old word (and when possible, vice versa) according to the University of South Florida free association normed database (Nelson et al., 2004).

In the perceptual task, the difference in brightness (Δb) between the two stimuli was calibrated for each subject and independently for each stimulus type. The brightness of a randomly located reference stimulus was fixed (middle gray). The brightness of the nonreference stimulus was titrated using a staircase procedure similar to previous experiments (Fleming et al., 2010, 2012b, 2014). During practice, we used a fixed, large step size two-down/one-up procedure until subjects reached 15 reversals or 90 trials. The step sizes followed recommended ratios to match the expected performance in memory blocks (García-Pérez, 1998). The experiment began with a Δb value determined by the average of the Δb values at each reversal, excluding the first one. Throughout the experiment, we kept a small step size staircase running to account for learning or tiredness.

A middle gray fixation cross subtending 0.3 degrees of visual angle was presented between the two stimuli on a black background. The reference stimulus in the perceptual task and all stimuli in the memory task were middle gray. All stimuli were surrounded by an isoluminant blue bounding box separated from the stimulus by a gap of at least 0.15 degrees of visual angle.

Behavioral data analysis

Data analysis was performed in MATLAB and statistical analysis in RStudio (R Studio Team, 2015). We estimated metacognitive efficiency by computing log(meta-d'/d') where d' is a signal detection theoretic measure of type I sensitivity and meta-d' is a measure of type II sensitivity (i.e., the degree to which a subject discriminates correct form incorrect responses) expressed in the same units as type I sensitivity (d') (Maniscalco and Lau, 2012; Fleming and Lau, 2014). Meta-d' indicates the d' that would have been predicted to give rise to the observed confidence rating data assuming a signal detection theoretic ideal observer. Meta-d' = d' indicates an optimal type II behavior for the observed type I behavior. Meta-d' greater or less than d' indicates metacognition that is better or worse, respectively, than the expected given task performance, as may occur, for instance, if first-order decisions and confidence are supported by partly parallel processing streams (Fleming and Daw, 2017). We used hierarchical Bayesian estimation to incorporate subject-level uncertainty in group-level parameter estimates (Fleming, 2017). Certainty on this parameter was determined by computing the 95% high-density interval (HDI) from the posterior samples (Kruschke, 2010). For correlation and individual differences analyses, we used single-subject Bayesian model fits. Two subjects were discarded for missing >10% of the trials (i.e., >1 SD from the average missed trials, which was 5%). Missed trials were not analyzed.

fMRI data acquisition

Brain images were acquired using a 3T Allegra scanner (Siemens). BOLD-sensitive functional images were acquired using a T2*-weighted gradient-echo echo-planar images (42 transverse slices, interleaved acquisition; TR, 2.34 s; TE, 30 ms; matrix size: 64 × 64; 3 × 3 mm in-plane resolution; slice thickness: 3 mm; flip angle: 90°; FOV: 126 mm). The main experiment consisted of three runs of 210 volumes and three runs of 296 volumes for the perceptual and memory tasks, respectively. We collected a T1-weighted MPRAGE anatomical scan (1 × 1×1 mm voxels; 176 slices) and local field maps for each subject.

fMRI data preprocessing

Imaging analysis was performed using SPM12 (Statistical Parametric Mapping; www.fil.ion.ucl.ac.uk/spm). The first five volumes of each run were discarded to allow for T1 stabilization. Functional images were realigned and unwarped using local field maps (Andersson et al., 2001) and then slice-time corrected (Sladky et al., 2011). Each participant's structural image was segmented into gray matter, white matter, CSF, bone, soft tissue, and air/background images using a nonlinear deformation field to map it onto template tissue probability maps (Ashburner and Friston, 2005). This mapping was applied to both structural and functional images to create normalized images to Montreal Neurological Institute (MNI) space. Normalized images were spatially smoothed using a Gaussian kernel (8 mm FWHM). We set a within-run 1 mm rotation and 4 mm affine motion cutoff criterion, which led to the exclusion of 4 subjects, leaving a total of 24 subjects whose functional and behavioral data were fully analyzed.

Univariate analysis

All of our general linear models (GLMs) focus on the "rating period" of each trial by specifying boxcar regressors beginning at the subjects' type I response and ending at their type II response (i.e., either confidence rating or number press). Motion correction parameters were entered as covariates of no interest along with a constant term per run. Regressors were convolved with a canonical hemodynamic response function. Low-frequency drifts were excluded with a 1/128 Hz high-pass filter. Missed trials were not modeled. For judgment-related (JR) analyses, we created a GLM with two regressors of interest per run to estimate BOLD response amplitudes in each voxel during the rating period in each trial of the confidence and follow blocks. For the confidence-level-related (CLR) parametric modulation analysis, a GLM was used to estimate BOLD responses in the confidence blocks. There were two regressors of interest in each run, one modeling the confidence rating period and another that encoded a parametric modulation by the four available confidence ratings (1–4).

Statistical inference. For the JR analysis, single-subject contrast images of the confidence and follow regressors were entered into a second-level random-effects analysis using one-sample t tests against zero to assess group-level significance. For the CLR parametric modulation analysis, single-subject contrast images of the parametric modulator were entered into a similar second-level random-effects analysis. For conjunction analyses of activations common to both domains, second-level maps thresholded at $p < 0.001$ (uncorrected) were intersected to reveal regions of shared statistically significant JR and CLR activity. Activations were visualized using MRIcro (http://www.mccauslandcenter.sc.edu/crnl/mricro). All second-level unthresholded statistical images were uploaded to Neurovault (Gorgolewski et al., 2015) (https://neurovault.org/collections/3232/).

ROI analysis. To define regions of interest (ROIs), 12 mm spheres were centered at MNI coordinates identified from previous literature (see Fig. 3*C*). ROIs in left rostrolateral PFC (L rlPFC) [−33, 44, 28], right rlPFC (R rlPFC) [27, 53, 25], and dorsal anterior cingulate cortex/presupplementary motor area (dACC/pre-SMA) [0, 17, 46] were created based on (Fleming et al., 2012b). The mask for precuneus (PCUN) [0, −64, 24] was based on (McCurdy et al., 2013). The MNI x-coordinates for the dACC/pre-SMA and PCUN masks were set to 0 to ensure bilaterality. Beta values were extracted from subjects' contrast images for the JR and CLR univariate analyses, respectively.

Multivoxel pattern analysis

Multivoxel pattern analysis (MVPA) was performed in MATLAB using the Decoding Toolbox (Hebart et al., 2014). We classified runwise beta images from GLMs modeling JR and CLR activity patterns in ROI and whole-brain searchlight analyses. ROI MVPAs were performed on normalized, smoothed images using the ROI spheres as masks. Previous work has shown that these preprocessing steps have minimal impact on support vector machine (SVM) classification accuracy while allowing meaningful comparison across subject-specific differences in anatomy, as in standard fMRI analyses (Kamitani and Sawahata, 2010; Op de Beeck, 2010). A single accuracy value per subject, per condition, and per ROI was

extracted and used for group analysis and statistical testing. As a control, we added a 6-mm-radius sphere centered at the ventricles [0 2 15].

Whole-brain searchlight analyses used 12 mm-radius spheres centered around a given voxel for all voxels on spatially realigned and slice-time-corrected images from each subject to create whole-brain accuracy maps. For group-level analyses, these individual searchlight maps were spatially normalized and smoothed using a Gaussian kernel (8 mm FWHM) and entered into one-sample *t* tests against chance accuracy (Hebart et al., 2014, 2016). Whole-brain cluster inference was performed in the same manner as in univariate analysis. Visualizations were made with Surf Ice (https://www.nitrc.org/projects/surfice/).

Before decoding, for JR activity pattern classification, we modeled two regressors of interest per run focused on the rating periods in the confidence and follow conditions. For classification of CLR activity patterns, we collapsed ratings 1 and 2 into a low-confidence regressor and ratings 3 and 4 into a high-confidence regressor to allow binary classification. The remaining parameters of no interest were specified as in the univariate case. For the CLR searchlight analysis, we used an exclusive mask of activity patterns associated with usage of the confidence scale obtained from the successful cross-classification of button presses (1–2 vs 3–4) between the confidence and follow conditions to eliminate low-level visuomotor confounds (see Fig. 4*D*).

In independent across-domain classifications, we used the runwise beta images reflecting JR and CLR activity as pattern vectors in a linear support vector classification model (as implemented in LIBSVM). We assigned each vector from each domain a label corresponding to the classes confidence (1) and follow (−1) in the JR analysis and low confidence (−1) and high confidence (1) in the CLR analysis. We trained an SVM with the vectors from one domain (three per class, six in total) and tested the decoder on the six vectors from the other domain (and vice versa) (see Fig. 4*A*, left), obtaining a mean average classification accuracy value for each of these two-way cross-classifications.

For within-domain classifications, we ran independent leave-one-run-out cross-validations for each domain on JR activity patterns (confidence vs follow) and CLR activity patterns (low vs high confidence). Pattern vectors from two of the three runs in each domain were used to train an SVM to predict the same classes in the vectors from the left-out run. We compared the true labels of the left-out run with the labels predicted by the model and iterated this process for the other two runs to calculate a mean cross-validated accuracy independently for each domain (see Fig. 4*A*, right).

We also tested the ability of confidence-related activity patterns to predict objective performance in the absence of confidence reports. We used a GLM that modeled low versus high confidence trials with a regressor that focused on the rating period and incorrect versus correct follow trials with a regressor that focused on the decision period (i.e., from stimulus onset to subjects' type I response). We performed a cross-classification analysis in which a decoder trained on confidence trials (low vs high confidence) was tested on pattern vectors from follow trials (incorrect vs correct) and vice versa (collapsed across domain). This confidence-objective performance generalization score was compared with a leave-one-run-out cross-validation analysis decoding low versus high confidence on confidence trials only (collapsed across domain). Together, these scores characterize whether a particular set of patterns are specific to confidence or also generalize to predict objective performance (Cortese et al., 2016) (see Fig. 5).

Individual differences

Metacognitive efficiency scores (log meta-*d*'/*d*') for each subject were estimated independently for the perceptual and memory tasks, together with a single score collapsed across domains. These scores were inserted as covariates in second-level analyses of within-perception, within-memory, and across-domain classifications of confidence-level-related activity, respectively, to assess the parametric relationship between metacognitive efficiency and decoding success.

Results

We analyzed the data from 24 subjects who underwent hemodynamic neuroimaging while performing two-alternative forced-choice discrimination tasks in perceptual and memory domains (Fig. 1*A*). In the perceptual task, subjects were asked to indicate the brighter of two stimuli (words or abstract shapes). In the memory task, subjects were asked to memorize exemplars of the same stimulus types and then select the previously learned stimulus from two stimuli presented on each trial. In half of the trials ("confidence" condition), subjects performed a metacognitive evaluation after the discrimination task by rating their confidence in the correctness of their response by selecting a number on a scale of 1 to 4 (1 = not confident; 4 = very confident). To differentiate metacognitive-related activity from visuomotor activity engaged by use of the confidence scale, in the other half of trials ("follow" condition), subjects were asked to respond according to a highlighted number without evaluating confidence in their response. To avoid stimulus-type confounds, two different types of stimuli, words and abstract shapes, were used in both tasks.

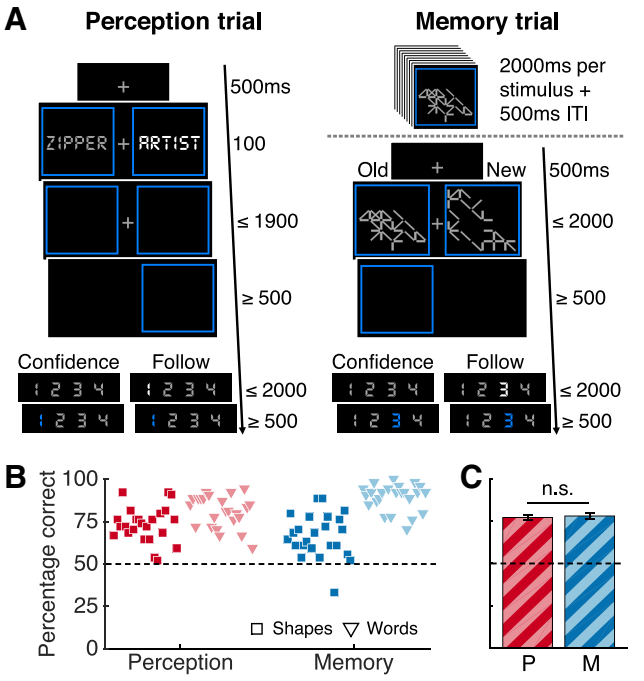

Figure 1. Task design and performance results. ***A***, Subjects performed two-alternative forced-choice discrimination tasks about perception and memory. In perception blocks, subjects selected the brighter of two stimuli. Memory blocks started with an encoding period and then subjects indicated in each trial which of two stimuli appeared during the encoding period. Abstract shapes and words were used as stimuli in both tasks. In confidence blocks, subjects rated their confidence and, in follow blocks, they pressed the highlighted number. ***B***, Percentage correct responses per block type in the confidence condition. Each marker represents a subject. ***C***, Mean percentage correct responses by domain averaged over subjects and stimulus types. Dotted lines indicate chance performance. Bars indicate SEM. n.s., Not significant; P, perception; M, memory.

Behavior

We first compared task performance, measured by percentage of correct responses, across condition, task, and stimulus type. A 2 × 2 × 2 repeated-measures ANOVA (confidence/follow × perception/memory × shapes/words) showed that performance was well matched across conditions (confidence vs follow) ($F_{(1,23)} = 3.036$, $p = 0.095$). None of the four paired *t* tests (domain × stimulus) comparing performance between the confidence and follow conditions returned a significant difference ($p > 0.05$). In the remainder of the behavioral analyses, we focused on the confi-

dence condition. Matching performance across stimulus type was more challenging because subjects' memory for words was expected to be considerably higher than that for abstract shapes trials based on pilot data (see Materials and Methods for details). Instead, we aimed to match subjects' performance independently for each stimulus type across task domains by titrating the difficulty of the perceptual task to approximate the performance expected for the corresponding stimulus type in the memory task (shapes: perceptual M = 73%, memory M = 67%; words: perceptual M = 81%, memory M = 89%; Fig. 1*B*). Critically, this ensured that performance was matched across task domains when averaging stimulus types across participants (perceptual: M = 77%, memory: M = 78%; paired *t* test $t_{(23)} = 0.38$, $p = 0.70$; Fig. 1*C*). A 2 × 2 repeated-measures ANOVA of performance in the confidence condition (perception/memory × shapes/words) confirmed there was no main effect of domain ($F_{(1,23)} = 0.15$, $p = 0.702$). However, we observed a main effect of stimulus type due to greater overall performance on words ($F_{(1,23)} = 75.69$, $p = 9.87 \times 10^{-9}$) and a domain × stimulus interaction due to a greater difference in performance between shapes and words in the memory compared with the perception task ($F_{(1,23)} = 16.74$, $p = 0.00045$).

Subjects were faster providing type I responses in perceptual trials (M = 636 ms) than in memory trials (M = 1222 ms). There was also a small difference in reaction times between shape (M = 967 ms) and word (M = 892 ms) trials. A 2 × 2 repeated-measures ANOVA confirmed a main effect of domain ($F_{(1,23)} = 367$, $p = 1.23 \times 10^{-15}$) driven by slower reaction types in the memory task. There was also a main effect of stimulus type on response time ($F_{(1,23)} = 8.95$, $p = 0.006$), as well as a significant domain × stimulus interaction due to a greater difference in reaction times between shapes and words in memory compared with the perception task ($F_{(1,23)} = 5.82$, $p = 0.024$).

As expected, subjects gave higher confidence ratings after correct decisions than after incorrect decisions (Fig. 2*A*) and mean confidence ratings were similar across task domains (perceptual M = 2.62, memory M = 2.47; paired *t* test $t_{(23)} = 1.26$, $p = 0.22$). Reaction times for confidence ratings were not different between domains (perceptual M = 518 ms, memory M = 516 ms; paired *t* test $t_{(23)} = 0.16$, $p = 0.87$). We next estimated log (meta-*d*'/*d*'), a metacognitive efficiency measure derived from signal detection theory that assays the degree to which confidence ratings distinguish between correct and incorrect trials (Maniscalco and Lau, 2012; Fleming and Lau, 2014; Fleming, 2017). We used hierarchical Bayesian estimation to incorporate subject-level uncertainty in group-level parameter estimates (Fleming, 2017). Metacognitive efficiency in the perceptual task was significantly lower than in the memory task ($p_{\theta>0} \sim 1$; for details, see Fig. 2*B* and the Materials and Methods), consistent with previous findings (Fleming et al., 2014). Metacognitive efficiency above optimality (meta-*d*' = *d*') in memory trials suggests subjects had better metacognition than expected given their task performance, whereas the suboptimal metacognitive efficiency in perceptual trials suggests that subjects had worse metacognition than expected given their task performance (assuming an ideal observer in both cases). We did not find a correlation between subjects' individual metacognitive efficiency scores in the perceptual and memory domains ($r_{(22)} = -0.076$; $p = 0.72$; Fig. 2*C*). We also evaluated the correlation coefficient within a hierarchical model of meta-*d*', which takes into account uncertainty in subject-level model fits (Fleming, 2017). The 95% confidence interval on the posterior correlation coefficient overlapped zero in this analysis ($\rho = 0.205$; HDI = [0.826, −0.358]), also indicating a dissociation between domains.

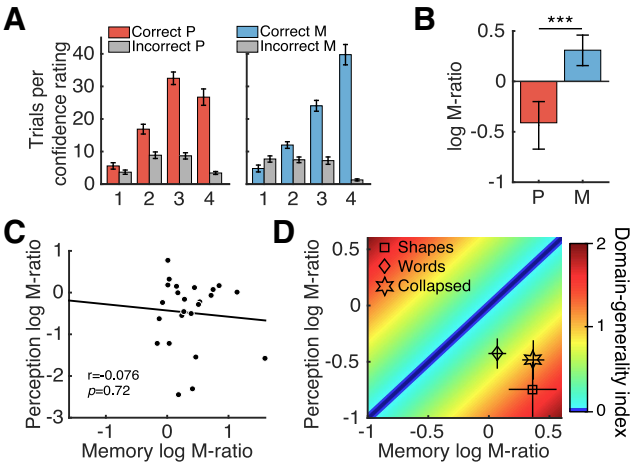

Figure 2. Metacognitive measures. ***A***, Mean number of correct and incorrect trials per confidence rating. ***B***, Metacognitive efficiency measured by log(meta-*d*'/*d*'). Zero indicates that metacognitive sensitivity (meta-*d*') is equal to task sensitivity *d*' (i.e., the *d*' that would have been predicted to give rise to the observed confidence rating data assuming a signal detection theoretic ideal observer). Group-level hierarchical Bayesian estimates differed significantly between domains. Error bars indicate 95% HDI from posterior samples. ***C***, Metacognitive efficiency scores obtained from single-subject Bayesian model fits were not correlated across perceptual and memory domains. ***D***, DGI for each subject that quantifies the similarity between their metacognitive efficiency scores in each domain (see main text for details). Greater DGI scores indicate less metacognitive consistency across domains. Mean log(meta-*d*'/*d*') values for each stimulus type in both domains are shown for reference. Bars in ***A*** and ***D*** indicate SEM. ***$p_{\theta>0} \sim 1$. P, Perception; M, memory.

We next estimated metacognitive efficiency separately for each stimulus type (Fig. 2*D*). A 2 × 2 repeated-measures ANOVA (perception/memory × shapes/words) indicated that metacognitive efficiency was greater for memory than perception ($F_{(1,23)} = 22.44$, $p = 8.97 \times 10^{-5}$). Importantly, there was no stimulus main effect ($F_{(1,23)} = 0.015$, $p = 0.902$) and there was no interaction between domain and stimulus type ($F_{(1,23)} = 2.835$, $p = 0.106$). To further assess a potential covariation between metacognitive abilities in each domain, we calculated for each subject a domain-generality index (DGI) that quantifies the similarity between scores in each domain for each participant (Fleming et al., 2014) as follows:

$$\mathrm{DGI} = |\log M_{\mathrm{P}} - \log M_{\mathrm{M}}|$$

where M_{P} = perceptual meta-*d*'/*d*' and M_{M} = memory meta-*d*'/*d*'. Lower DGI scores indicate more similar metacognitive efficiencies between domains (DGI = 0 indicates identical scores). Mean DGI for shapes (1.42), words (0.66), and collapsed by stimulus type (0.95) were higher than zero (Fig. 2*D*). Metacognition for words was behaviorally more stable across domains because the DGI was smaller than for shapes (paired *t* test: $t_{(23)} = 2.86$; $p = 0.009$). Together, these results suggest domain-specific constraints on metacognitive ability.

fMRI analyses

We next turned to our fMRI data to assess the overlap between neural substrates engaged when metacognitive judgments are made during perceptual and memory tasks. A full understanding of the neural substrates of metacognition requires an independent examination of the process of engaging in a metacognitive task and the level of confidence expressed by the subject (Chua et

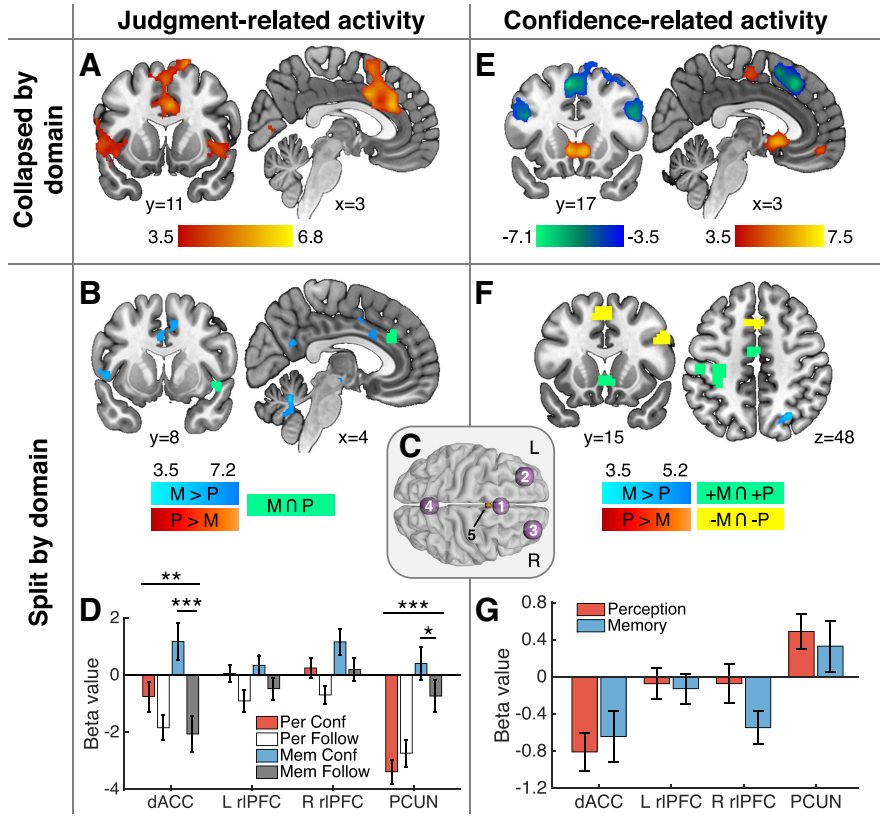

Figure 3. fMRI univariate analysis results. ***A–D***, JR activity. ***A***, Whole-brain analysis of significant activation in the confidence > follow contrast (collapsed by domain); there were no significant clusters in the follow > confidence contrast. ***B***, (Memory confidence > memory follow) > (perception confidence > perception follow) interaction contrast (blue). There were no significant clusters in the reverse contrast. The conjunction of memory confidence > memory follow and perception confidence > perception follow contrasts is indicated in green. ***C***, Spherical binary masks of four a priori ROIs, 1 = dACC/pre-SMA, 2 = L rlPFC, 3 = R rlPFC, and 4 = PCUN, and an ROI in the ventricles (5) used as a control region in multivariate analyses (Fig. 4). ***D***, Estimated mean beta values for JR activity by domain in the main four ROIs displayed in ***C***. ***E–G***, CLR activity. ***E***, Whole-brain analysis of activity parametrically modulated by level of confidence (collapsed by domain). Hot colors indicate a positive correlation with confidence and cool colors a negative correlation. ***F***, Memory > perception contrast (blue) testing for differences between the parametric effect of confidence by domain; there were no significant clusters in the perception > memory contrast. A conjunction analysis revealed shared activity that was positively (green) and negatively (yellow) correlated with confidence levels in both domains. ***G***, Estimated mean beta values of CLR activity in the main four ROIs displayed in ***C***. All displayed whole-brain activations are significant at a cluster-defining threshold $p < 0.001$, corrected for multiple comparisons. $p_{FWE} < 0.05$; except for conjunction analyses, in which we computed the intersection of two independent maps thresholded at $p < 0.001$, uncorrected. Images are displayed at $p < 0.001$. Graded color bars reflect T-statistics. Error bars indicate SEM. ***$p < 0.001$; **$p < 0.01$; *$p < 0.05$. L, Left; R, right; P, perception; M, memory.

al., 2014). To this end, in both univariate and multivariate analyses, we focused on two distinct features of metacognition-related activity. First, we assessed brain regions engaged in JR activity (i.e., the difference between confidence trials requiring a metacognitive judgment and the follow condition). Second, we assessed brain regions engaged in CLR activity. In univariate CLR analysis, we focused on the parametric relationship between confidence ratings (1–4) and neural activity. In multivariate CLR analysis, we collapsed ratings 1 and 2 into a low-confidence category and ratings 3 and 4 into a high-confidence category to allow binary classification of activity patterns.

Univariate results

JR activity. In standard univariate analyses, we found elevated activity in dACC/pre-SMA, bilateral insulae, and superior and middle frontal gyri when contrasting the confidence against the follow condition (collapsed by domain), which is consistent with previous findings (Fleming et al., 2012b) (Fig. 3*A*). There were no significant clusters of activity in the reverse contrast (follow > confidence). Splitting the data by domain (Table 1), an interaction contrast (memory confidence > memory follow) > (perception confidence > perception follow) revealed significant clusters of activity in middle cingulate gyrus, left insula, PCUN, left hippocampus, and cerebellum (Fig. 3*B*, blue). No significant clusters of activity were found in the reverse interaction contrast. In a conjunction analysis, elevated activity for the confidence > follow condition was observed across both perception and memory trials in anterior cingulate and right insula (Fig. 3*B*, green).

To further quantify these effects for each task domain, we focused on four a priori ROIs in the dACC/pre-SMA, bilateral rlPFC, and PCUN (Fig. 3*C* and the Materials and Methods), which previous studies have found to be recruited by perceptual and memory metacognition (Fleck et al., 2006; Fleming et al., 2010, 2012b; Baird et al., 2013; McCurdy et al., 2013). In a series of repeated-measures 2 × 2 ANOVAS (condition × task), we found a main effect of greater activity on confidence compared with follow trials in all ROIs except PCUN, where instead we observed a main effect of task, with increased activity on memory trials (Fig. 3*D*; condition main effect: dACC/pre-SMA, $F_{(1,23)} = 19.34$, $p = 0.0002$; left rlPFC, $F_{(1,23)} = 6.62$, $p = 0.017$; right rlPFC, $F_{(1,23)} = 9.28$, $p = 0.006$; PCUN: $F_{(1,23)} = 0.40$, $p = 0.532$; task main effect: dACC/pre-SMA, $F_{(1,23)} = 2.33$, $p = 0.14$; left rlPFC, $F_{(1,23)} = 0.95$, $p = 0.34$; right rlPFC, $F_{(1,23)} = 4.94$, $p = 0.036$; PCUN: $F_{(1,23)} = 36.78$, $p = 3.47 \times 10^{-6}$). We found that the difference between confidence and follow trials was greater in memory than in perception trials in dACC/pre-SMA, recapitulating the whole-brain results (condition × task interaction; $F_{(1,23)} = 12.16$, $p = 0.0019$; paired t test, memory: $t_{(23)} = 5.47$, $p = 0.0001$, perceptual: $t_{(23)} = 1.92$, $p = 0.067$). A similar interaction pattern was observed in the PCUN (condition × task interaction: $F_{(1,23)} = 15.86$, $p = 0.0006$; paired t test, memory: $t_{(23)} = 2.43$, $p = 0.023$, perceptual: $t_{(23)} = -1.54$, $p = 0.136$). There were no interactions in frontal regions (left rlPFC, $F_{(1,23)} = 0.07$, $p = 0.795$; right rlPFC, $F_{(1,23)} = 0.002$, $p = 0.968$). These results are compatible with previous findings indicating a distinctive contribution of PCUN to memory metacognition (Baird et al., 2013; McCurdy et al., 2013).

CLR activity. We next sought to investigate the parametric relationship between confidence level and neural activity. Collapsing across domains, we found activity in the left precentral and postcentral gyri, the posterior midline, ventral striatum, and ventromedial PFC (vmPFC) correlated positively with confidence ratings (Fig. 3*E*, hot colors). We also replicated negative correlations between confidence and activation in dACC/pre-SMA, parietal cortex, and bilateral PFC that have been reported in several previous studies (Fleck et al., 2006; Fleming et al.,

Table 1. Univariate fMRI analysis: judgment-related activity interacted with domain

Contrast	Label	Voxels at $p < 0.001$	p_{fwe} cluster-corrected	Peak *z*-score	Peak voxel MNI coordinates	Laterality
Memory (C > F) > perception (C > F)	Cerebellum	70	0.015	5.17	3, −58, −25	L/R
	Insula	109	0.001	4.89	−33, −10, −7	L
	Posterior cingulate, precuneus	84	0.006	4.29	3, −58, 23	L/R
	Postcentral gyrus, BA3	99	0.002	4.25	−45, −31, 62	L
	Hippocampus, parahippocampal gyrus, fusiform area	100	0.002	4.13	−21, −37, −16	L
	Thalamus	66	0.019	4.10	9, −25, −7	R
	Middle and anterior cingulate gyrus, SMA, BA24, BA32	194	<0.001	4.09	−6, 5, 32	L/R
	Inferior frontal gyrus, BA47	61	0.026	4.05	−42, 20, −4	L
Conjunction memory (C > F) ∩ perception (C > F)	Cingulate gyrus, BA32	12			6, 32, 29	R
	Insula	7			45, 11, −7	R

Significant activations at cluster-defining threshold $p < 0.001$, corrected for multiple comparisons at $p_{fwe} < 0.05$. Conjunction of significant activations at cluster-defining threshold $p < 0.001$, uncorrected, of memory (C > F) and Perception (C > F) contrasts. For more information, see Figure 3*B*. C, Confidence; F, follow.

Table 2. Univariate fMRI analysis: confidence-level-related activity interacted with domain

Contrast	Label	Voxels at $p < 0.001$	p_{fwe} cluster-corrected	Peak *z*-score	Peak voxel MNI coordinates	Laterality
Memory > perception	Precuneus, BA7	93	0.003	4.21	33, −70, 20	R
Conjunction (+M ∩ +P)	Precentral and postcentral gyri, BA6, BA4, BA3	167			−30, −25, 44	L
	Postcentral gyrus	27			−27, −46, 59	L
	SMA, BA6	21			−3, −10, 50	L/R
	Ventral striatum	16			3, 11, 7	L/R
	Cerebellum	7			9, −58, −13	R
	Postcentral gyrus	5			−48, −22, 44	L
Conjunction (−M ∩ −P)	Middle frontal gyrus	29			45, 26, 20	R
	Pre-SMA, BA8	19			0, 14, 50	L/R

Significant activations at cluster-defining threshold $p < 0.001$, corrected for multiple comparisons at $p_{fwe} < 0.05$. Conjunction of significant activations at cluster-defining threshold $p < 0.001$, uncorrected, of memory (M) and perception (P) contrasts of positive and negative correlations with confidence level. For more information, see Figure 3*F*.

2012b; Baird et al., 2013; Hebart et al., 2016) (Fig. 3*E*, cool colors). When testing for differences between these parametric regressors by domain (Table 2), a memory > perception contrast revealed a significant cluster of activity in right parietal cortex (Fig. 3*F*, blue), whereas there was no significant activity in a perception > memory contrast. Shared positive correlations between confidence and activity in perception and memory trials were found in ventral striatum and in left precentral and postcentral gyri, the latter consistent with use of the right hand to provide confidence ratings (conjunction analysis; Fig. 3*F*, green). Shared negative correlations with confidence were found in regions of right dorsolateral PFC and medial PFC, overlapping with pre-SMA (Fig. 3*F*, yellow).

Complementing the ROI analysis of JR activity, we performed an ROI analysis of CLR activity that recapitulated the whole-brain results. We observed negative relationships between confidence and activity in dACC/pre-SMA and positive relationships in PCUN. Importantly, no significant differences in the parametric effect of confidence were found between domains in any of our a priori ROIs (Fig. 3*G*; paired *t* tests: dACC/pre-SMA, $t_{(23)} = -0.47$, $p = 0.643$; left rlPFC, $t_{(23)} = 0.23$, $p = 0.820$; right rlPFC, $t_{(23)} = 1.62$, $p = 0.119$; PCUN: $t_{(23)} = 0.56$, $p = 0.583$). Together with the lack of marked domain-specific differences in confidence-related activity at the whole-brain level, these results are suggestive of an absence of domain specificity in confidence-related activity. However, a lack of difference between univariate activation profiles is not necessarily conclusive. For instance, differences in confidence level may be encoded in fine-grained spatial patterns of activity even when the overall BOLD activity is evenly distributed across confidence levels (Cortese et al., 2016). Similarly, whereas metacognition-related activity may show similar overall levels of activation across tasks, distributed activity patterns in frontal and parietal areas may carry distinct task-specific information (Hebart et al., 2016; Cole et al., 2016). We next turned to multivariate analysis methods, which are sensitive to differences in spatial activity patterns, to test this hypothesis.

Multivariate results

We performed a series of MVPAs (Fig. 4*A*) focused on both JR activity patterns and CLR activity patterns.

ROI analysis of JR activity patterns. If metacognitive judgments are based on domain-general processes (i.e., shared across perceptual and memory tasks), then a decoder trained to classify JR activity patterns in perceptual trials should accurately discriminate JR activity patterns when tested on memory trials (and vice versa). Alternatively, domain-specific activity profiles would be indicated by significant within-domain classification of JR activity patterns in the absence of across-domain transfer. To adjudicate between these hypotheses, we performed an SVM decoding analysis using as input vectors the runwise beta images pertaining to confidence and follow trials obtained from a GLM (12 input vectors in total). For within-domain classification, we used standard leave-one-out independent cross-validations for each domain and we tested for across-domain generalization using a cross-classification analysis (see Materials and Methods for details). Chance classification in both analyses was 50%.

Mean within-domain classification results were significantly above chance in all ROIs (one-sample *t* tests Bonferroni corrected for multiple comparisons $\alpha = 0.05/4 = 0.0125$: dACC/pre-SMA $t_{(23)} = 5.77$, $p = 6.99 \times 10^{-6}$; L rlPFC $t_{(23)} = 3.27$, $p = 0.003$; R rlPFC $t_{(23)} = 4.47$, $p = 0.0002$; PCUN $t_{(23)} = 2.98$, $p = 0.007$; Fig. 4*B*, red and blue striped bars). In contrast, across-domain generalizations were not significantly different from chance in any ROI (one-sample *t* test Bonferroni corrected: dACC/pre-SMA, $t_{(23)} = 1.95$, $p = 0.06$; L rlPFC, $t_{(23)} = 0.79$, $p = 0.44$; R rlPFC, $t_{(23)} = 1.24$, $p = 0.23$; PCUN $t_{(23)} = 1.40$, $p = 0.17$; Fig. 4*B*, yellow bars).

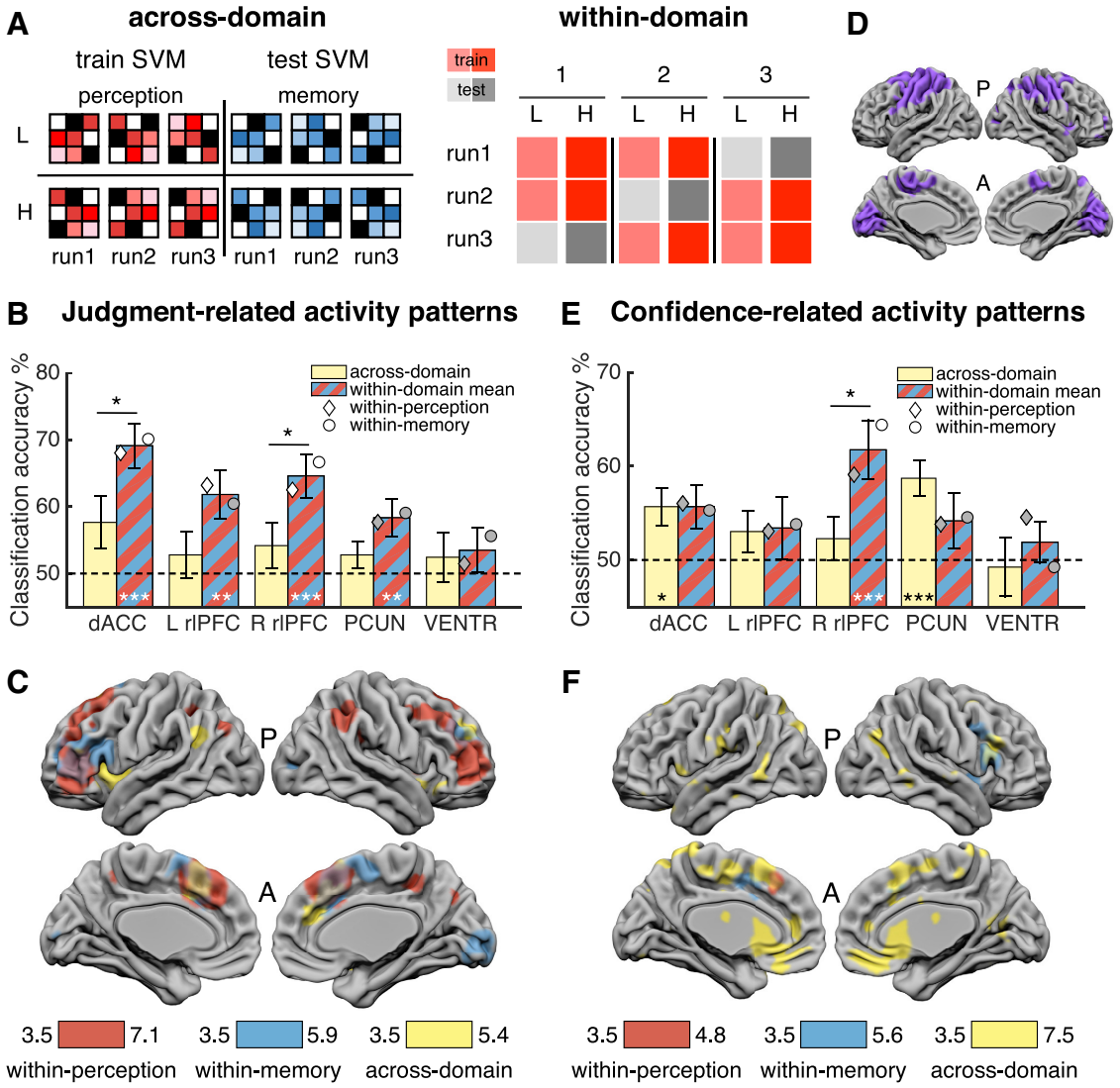

Figure 4. MVPA results. Classification designs. ***A***, Left, Across-domain classification design. Pattern vectors (runwise beta images) from one domain were used to train an SVM decoder on two classes and then tested in a cross-classification of the same two classes using vectors from the other domain (and vice versa). Classification of low (L) and high (H) confidence levels is illustrated. Right, Within-domain classification design. Pattern vectors of two classes (e.g., low and high confidence) pertaining to one domain were used to train a decoder in a leave-one-run-out design that was then tested in the left-out pair. The process was iterated three times to test pairs from every run. An identical, independent cross-validation was performed on vectors from the other domain. ***B***, ***C***, JR activity patterns. ***B***, ROI results for across-domain (yellow) and mean within-domain (red-blue stripe) classification accuracy of Confidence vs Follow trials. ***C***, Searchlight analysis for same classifications in ***B***. ***D***, Low-level visuomotor mask used in ***F*** (see main text and Materials and Methods for details). ***E***, ***F***, CLR activity patterns. ***E***, Low versus high confidence classification accuracy results. ***F***, Searchlight analysis for the same classifications in ***E*** exclusively masked for visuomotor-related activity patterns. Bars in ***B*** and ***E*** indicate means and error bars indicate SEM. Dashed lines indicate chance classification (50%). Diamonds and circles indicate mean independent classification in perception and memory trials, respectively. White diamonds/circles indicate classification was significantly different from chance, Bonferroni corrected. All clusters in ***C*** and ***F*** are significant at cluster-defining threshold $p < 0.001$, corrected for multiple comparisons at $p_{FWE} < 0.05$. Image is displayed at $p < 0.001$. Color bars indicate T-scores. A, anterior; P, posterior. ***$p \leq 0.001$; **$p \leq 0.01$; *$p < 0.05$; all one-sample t tests are Bonferroni corrected.

As a control, classification accuracy in the ventricles was not different from chance (across: $t_{(23)} = 0.66$, $p = 0.52$; within: $t_{(23)} = 1.04$, $p = 0.31$). This suggests that the patterns of activity that distinguish metacognitive judgments from the visuomotor control condition in one domain are distinct from analogous patterns in the other domain. In particular, within-domain classification accuracy was significantly different from across-domain classification accuracy in (dACC/pre-SMA: $t_{(23)} = 2.88$, $p = 0.008$) and right rlPFC ($t_{(23)} = 2.24$, $p = 0.035$). These results are consistent with the hypothesis that metacognitive judgments recruit domain-specific patterns of cortical activity in PFC.

Searchlight analysis of JR activity patterns. We ran a similar decoding analysis using an exploratory whole-brain searchlight, obtaining a classification accuracy value per voxel (Hebart et al., 2014). Consistent with our ROI results, we observed significant within-domain classification in large swathes of bilateral PFC for both perception (red) and memory (blue) (Fig. 4*C*, Table 3). Within-perception classification was also successful in parietal regions, the PCUN in particular, and within-memory activity patterns were classified accurately in occipital regions. We also identified clusters showing significant across-domain generalization (yellow) in dACC, pre-SMA, SFG (BA9), supramarginal gyrus (BA40), and bilateral IFG/insula, consistent with univariate results (Fig. 3*A*).

ROI analysis of CLR activity patterns. We next investigated whether confidence is encoded in a domain-general or domain-specific fashion by applying a similar approach to discriminate low versus high confidence trials. In this case, ROI univariate analyses did not reveal any differences in confidence-related ac-

Table 3. Judgment-related activity obtained from whole-brain searchlight classification analyses

Classification	Label	Voxels at $p < 0.001$	p_{FWE} cluster-corrected	Peak *z*-score	Peak voxel MNI coordinates	Laterality
Across-domain	Insula/IFG	592	0.001	4.31	−33, 23, 5	L
	SFG and BA9	928	<0.001	4.29	24, 53, 38	R
	dACC/pre-SMA			4.19	−12, 14, 41	L
				4.01	12, 44, 23	R
	Supramarginal gyrus, BA40	207	0.039	3.93	−51, −55, 29	L
	IFG/insula/STG	275	0.016	3.61	48, 5, −13	R
Within-perception	Superior, middle, medial, inferior FG, dACC, pre-SMA, BA8, BA10, BA32	4699	<0.001	5.10	−12, 14, 59	L/R
	Parietal cortex, precuneus, supramarginal gyrus, BA7, BA40	1141	<0.001	4.30	51, −46, 44	R
	Precuneus, BA40	339	0.010	4.21	−36, −46, 47	L
Within-memory	Middle and superior FG, BA9, BA10	325	0.008	4.54	39, 41, 38	R
	Lingual gyrus, cuneus, calcarine	510	0.001	4.53	9, −91, 2	R
	Superior FG, dACC, pre-SMA	1066	<0.001	4.33	6, −4, 62	L/R
	Middle, inferior FG, BA9, BA10, BA45, BA46	1003	<0.001	4.09	−42, 38, 2	L

Significant activations at cluster-defining threshold $p < 0.001$, corrected for multiple comparisons at $p_{FWE} < 0.05$. For more information, see Figure 4*C*.

tivity between domains (Fig. 3*G*). We hypothesized that if confidence level is encoded by domain-general neural activity patterns, then it should be possible to train a decoder to discriminate low (1–2) from high (3–4) confidence rating patterns in the perceptual task and then accurately classify confidence patterns on the memory task (and vice versa). In the absence of across-domain classification, significant within-domain classification is indicative of CLR domain-specific activity patterns. ROI cross-classifications and cross-validations were performed in a similar fashion as above (Fig. 4*A*). Two subjects did not provide ratings for one of the classes in at least one run and were left out from the main analysis to avoid entering unbalanced training data into the classifier; however, including those subjects did not change the main result.

One-sample *t* tests (Bonferroni-corrected) showed that across-domain classification of confidence was significantly above chance in dACC/pre-SMA ($t_{(21)} = 2.83$, $p = 0.010$) and PCUN ($t_{(21)} = 4.69$, $p = 0.0001$), indicative of a generic confidence signal, but not in rlPFC (left: $t_{(21)} = 1.36$, $p = 0.19$; right: $t_{(21)} = 0.97$, $p = 0.34$) (Fig. 4*E*, yellow bars). In contrast, mean within-domain classification accuracy was significantly above chance in right rlPFC ($t_{(21)} = 3.75$, $p = 0.001$), but not in the other ROIs (dACC/pre-SMA: $t_{(21)} = 2.42$, $p = 0.025$; left rlPFC: $t_{(21)} = 1.03$, $p = 0.32$; PCUN: $t_{(21)} = 1.42$, $p = 0.17$; Bonferroni-corrected; Fig. 4*E*, red and blue striped bars). The mean confidence classification accuracy in right rlPFC was 62% (perceptual = 59%, memory = 64%), notably above a recently estimated median 55% for decoding task-relevant information in frontal regions (Bhandari et al., 2017). Importantly, classification accuracy in this ROI also differed from the corresponding across-domain classification accuracy (paired *t* test $t_{(21)} = 2.37$, $p = 0.028$). Classification accuracy in the ventricles was not different from chance (across: $t_{(21)} = -0.24$, $p = 0.81$; within: $t_{(21)} = 0.86$, $p = 0.40$). When subjects with unbalanced data were included, within-domain classification accuracy in right rlPFC remained at 62%, significantly above chance ($t_{(23)} = 4.22$; $p = 0.0003$) and significantly different from across-domain classification accuracy (paired *t* test: $t_{(23)} = 2.54$; $p = 0.018$). Together, these results suggest the coexistence of two kinds of CLR neural activity: dACC/pre-SMA and PCUN encode a generic confidence signal, whereas patterns of activity in right rlPFC were modulated by task, allowing within- but not across-domain classification of confidence level.

Searchlight analysis of CLR activity patterns. We ran a similar decoding analysis of confidence level using an exploratory whole-brain searchlight, obtaining a classification accuracy value per voxel. Here, we leveraged the follow trials as a control for low-level visuomotor confounds by exclusively masking out activity patterns associated with usage of the confidence scale (Fig. 4*D*). The remaining activity patterns can therefore be ascribed to CLR signals that do not encode visual or motor features of the rating (Fig. 4*F*, Table 4). We found widespread across-domain classification of confidence (yellow) in a predominantly midline network including a large cluster encompassing dACC/pre-SMA, vmPFC, and ventral striatum. Domain-specific CLR activity patterns were successfully decoded from right PFC (insula, IFG, BA9, BA46) in memory trials (blue) and were also independently decoded in both domains from dACC/pre-SMA.

Generalization of CLR activity to objective performance. To further address the question of how confidence judgments may relate to activity patterns, we examined the relationship between objective task accuracy and confidence. Previous work suggested that the neural basis (and associated activation patterns) of confidence and performance may be partly distinct (Rounis et al., 2010; Cortese et al., 2016). Specifically, we tested the hypothesis that we could train a decoder using CLR activity patterns to classify objective performance-related activity patterns (correct/incorrect) on follow trials (and vice versa) in a cross-classification analysis (collapsed across domain). This analysis confirmed that activity patterns in dACC/pre-SMA ($t_{(21)} = 2.38$, $p = 0.027$) and right rlPFC ($t_{(21)} = 2.64$, $p = 0.015$) could predict objective accuracy levels in follow trials above chance (Fig. 5, light gray; uncorrected), but not in left rlPFC ($t_{(21)} = 1.49$, $p = 0.15$) or PCUN ($t_{(21)} = -0.46$, $p = 0.65$). We then compared these decoding scores with a leave-one-run-out cross-validation decoding analysis of low versus high confidence on confidence trials only (collapsed by domain; Fig. 5, dark gray; uncorrected). Consistent with the analyses reported in Figure 4*E* (yellow), this decoder was unable to classify domain-general confidence patterns of activity in right rlPFC ($t_{(21)} = 1.00$; $p = 0.33$), but was above chance in dACC/pre-SMA ($t_{(21)} = 3.80$, $p = 0.001$), left rlPFC ($t_{(21)} = 2.26$, $p = 0.034$), and PCUN ($t_{(21)} = 2.56$, $p = 0.018$). Critically, in PCUN, confidence classification was significantly greater than confidence-performance generalization, which was at chance (paired *t* test $t_{(21)} = 2.16$, $p = 0.043$). This result indicates that confidence-related patterns in PCUN do not generalize to predict objective performance, consistent with a partly distinct coding of information relevant to task performance and confidence. In contrast, in dACC/pre-SMA, general confidence level and performance could be predicted from common patterns of activation.

Table 4. Confidence level-related activity obtained from whole-brain searchlight classification analyses

Classification	Label	Voxels at $p < 0.001$	p_{FWE} cluster-corrected	Peak *z*-score	Peak voxel MNI coordinates	Laterality
Across-domain	Inferior FG, ventral striatum, ACC, BA11, BA32	2434	<0.001	5.20	−21, 44, 2	L/R
	Superior and middle temporal gyri	176	<0.001	4.99	−57, −52, 20	L
	Middle temporal gyrus, anterior cerebellum	1193	<0.001	4.92	3, −46, −34	R
	Middle cingulate gyrus	46	0.017	4.85	−6, −10, 47	L/R
	Superior temporal gyrus	142	<0.001	4.68	66, −7, 5	R
	Precuneus, BA7, BA19	543	<0.001	4.68	−12, −61, 65	L/R
	Superior temporal gyrus	110	<0.001	4.68	−42, 26, −28	L
	Posterior cerebellum	96	<0.001	4.54	−9, −85, −22	L
	SMA, BA6	84	0.001	4.53	−9, −10, 62	L/R
	Superior FG, dACC, pre-SMA	482	<0.001	4.39	−3, 11, 56	L/R
	Postcentral gyrus	186	<0.001	4.33	−57, −19, 23	L
	Middle FG	143	<0.001	4.28	36, 5, 26	R
	Fusiform and parahippocampal gyri	63	0.003	4.27	36, −19, −34	R
	Middle cingulate cortex	64	0.003	4.25	0, −31, 53	L/R
	Posterior cerebellum	53	0.008	4.10	−48, −58, −37	L
	Inferior FG	38	0.039	3.84	27, 23, −25	R
Within-perception	Superior FG, dACC/pre-SMA	185	0.032	3.88	−6, 29, 47	L/R
Within-memory	Inferior and middle FG, insula	503	<0.001	4.35	57, 23, 14	R
	dACC, pre-SMA	211	0.021	3.79	−9, 2, 47	L/R

Accuracy maps were masked to exclude visuomotor-related activity (see Fig. 4*D*). Significant activations at cluster-defining threshold $p < 0.001$, corrected for multiple comparisons at $p_{FWE} < 0.05$. For more information, see Figure 4*F*.

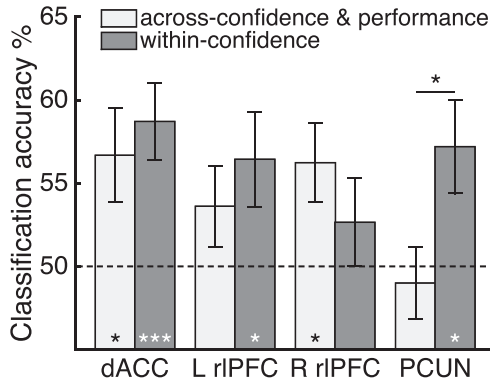

Figure 5. Generalization of confidence-related activity to objective accuracy. Light gray bars denote mean cross-classification accuracy results obtained from training a decoder on CLR activity and testing it on objective accuracy (correct/incorrect) activity patterns in the follow condition (and vice versa). Dark gray bars denote decoding accuracy for a leave-one-out cross-validation of low and high confidence on confidence trials only (collapsed by domain). Bars indicate group means and SEM. Dotted line indicates chance level. $*p < 0.05$; $***p < 0.001$.

Metacognitive efficiency and CLR activity classification. Finally, we reasoned that, if confidence-related patterns of activation are contributing to metacognitive judgments, then they may also track individual differences in metacognitive efficiency. To test for such a relation, we investigated whether individual metacognitive efficiency scores collapsed across domains and independently in each domain predicted searchlight classification accuracy of confidence level. We did not find any significant clusters after whole-brain correction for multiple comparisons in domain-general, within-perception, or within-memory analyses. However, memory metacognitive efficiency predicted memory confidence classification accuracy in a cluster in right PCUN and left precentral gyrus ($p < 0.001$, uncorrected), whereas perceptual metacognitive efficiency predicted perceptual confidence classification accuracy in left middle frontal gyrus, right vmPFC, bilateral temporal gyri, and cerebellum ($p < 0.001$, uncorrected). Previous studies (Fleming et al., 2010, 2014; McCurdy et al., 2013) have reported similar relations between perceptual and memory metacognitive efficiency and individual differences in the structure of prefrontal and parietal cortex, respectively. Although we do not interpret these findings further here, for completeness, second-level unthresholded statistical images of these analyses were uploaded to Neurovault to inform future work (Gorgolewski et al., 2015) (https://neurovault.org/collections/3232/).

Discussion

When performing a cognitive task, confidence estimates allow for comparisons of performance across a range of different scenarios (de Gardelle and Mamassian, 2014). Such estimates must also carry information about the task context if they are to be used in decision making. Here, we investigated the domain generality and domain specificity of representations that support metacognition of perception and memory.

Unlike previous studies (McCurdy et al., 2013), subjects' performance was matched between domains for two different types of stimulus, thereby eliminating potential performance and stimulus confounds. Subjects' confidence ratings were also matched between domains and followed expected patterns of higher ratings after correct decisions than after incorrect decisions. Metacognitive efficiency scores between tasks were not correlated and metacognitive efficiency scores in the memory task were superior to those in the perceptual task. Using univariate and multivariate analyses, we showed the existence of both domain-specific and domain-general metacognition-related activity during perceptual and memory tasks. We report four main findings and discuss each of these in turn.

First, we obtained convergent evidence from both univariate and multivariate analyses that a cingulo-opercular network centered on dACC/pre-SMA encodes a generic signal predictive of confidence level and objective accuracy across memory and perceptual tasks. Previous studies of metacognition have implicated the cingulo-opercular network in tracking confidence level (Fleck et al., 2006; Fleming et al., 2012b; Hilgenstock et al., 2014; Hebart et al., 2016). However, we go beyond these previous studies to provide evidence that these signals generalize to predict confidence across two distinct cognitive domains. This finding is consistent with posterior medial frontal cortex as a nexus for monitoring the fidelity of generic sensorimotor mappings, building on previous findings that error-related event-related potentials originating from this region are sensitive to variation in subjective certainty (Scheffers and Coles, 2000; Boldt and Yeung, 2015). The activity

in dACC/pre-SMA was also consistently elevated by the requirement for a metacognitive judgment (Fleming et al., 2012b). However, the results regarding the generalizability of the pattern of these increases across tasks were inconclusive. Whole-brain searchlight analysis revealed successful cross-classification of these activity patterns in dACC and insular regions, consistent with the results of the univariate analysis. These patterns of activity, however, did not generalize across tasks in a predefined region of interest centered in dACC/pre-SMA.

Although both dACC/pre-SMA and PCUN showed significant domain-general decoding of confidence, in PCUN these patterns did not generalize to also predict changes in objective accuracy. Whereas performance and subjective confidence may both depend on similar decision variables (Kiani and Shadlen, 2009; Fleming and Daw, 2017), behavioral dissociations between these quantities are also consistent with distinct internal states contributing to decisions and confidence ratings (Busey et al., 2000; Fleming and Daw, 2017). For instance, hierarchical models of confidence formation suggest a downstream network "reads out" decision-related information in a distinct neural population (Insabato et al., 2010). The observed lack of cross-classification in PCUN (Fig. 5) is consistent with the recent observation of distinct neural patterns of activity pertaining to confidence and first-order performance revealed through multivoxel neurofeedback in frontal and parietal regions (Cortese et al., 2016).

Second, in lateral anterior frontal cortex, we found activity patterns that tracked both the requirement for metacognitive judgments and level of confidence. Large swathes of lateral PFC distinguished activity patterns pertaining to metacognitive judgments that were specific for each domain. Critically, however, confidence-related activity patterns were selective for domain in right rlPFC (Fig. 4*E*): they differed according to whether the subject was engaged in rating confidence about perception or memory. Such signals may support the "tagging" of confidence with contextual information, thereby facilitating the use of confidence for behavioral control (Donoso et al., 2014; Purcell and Kiani, 2016). The identity of perceptual and memory tasks can be reliably decoded from activity in right PFC neural populations (Mendoza-Halliday and Martinez-Trujillo, 2017), consistent with the possibility that this contextual information is recruited during confidence rating. It is possible that anterior prefrontal regions combine generic confidence signals with domain-specific information to fine-tune decision making and action selection in situations in which subjects need to regularly switch between tasks or strategies on the basis of their reliability (Donoso et al., 2014). An alternative hypothesis, also compatible with our data, is that PFC first estimates the confidence level specifically for the current task, which is then relayed to medial areas to recruit the appropriate resources for cognitive control in a task-independent manner. Processing dynamics may also unfold simultaneously in both areas. These possibilities echo a longstanding debate in the cognitive control literature on the relative primacy of medial and lateral PFC in the hierarchy of control (Kerns et al., 2004; Tang et al., 2016). Further inquiry and development of computational models of the hierarchical or parallel functional coupling of these networks in metacognition is necessary.

Third, we obtained convergent evidence that PCUN plays a specific role in metamemory judgments. In univariate fMRI analyses, we found that the requirement for a metacognitive judgment recruited our preestablished region of interest centered on PCUN only on memory, but not perceptual, trials (Fig. 3*D*). Individual metacognitive efficiency scores in memory trials predicted classification accuracy in a more dorsal precuneal region, whereas individual differences in metacognitive efficiency scores in perceptual trials predicted classification accuracy in vmPFC (albeit at uncorrected thresholds). These findings are consistent with the medial parietal cortex making a disproportional contribution to memory metacognition (Simons et al., 2010; Baird et al., 2013; McCurdy et al., 2013) and offer a potential explanation for a decrease in perceptual, but not memory, metacognitive efficiency seen in patients with frontal lesions (Fleming et al., 2014). However, we do not conclude that PCUN involvement is specific to metamemory. We note that univariate negative correlations with confidence were found also on perceptual trials and multivariate classification results in PCUN indicated the presence of both perceptual- and memory-related signals. This dual involvement of the PCUN in perception and memory metacognition is consistent with previous studies suggesting a relationship between PCUN structure and visual perceptual metacognition (Fleming et al., 2010; McCurdy et al., 2013).

Fourth, we found in both univariate and multivariate whole-brain analyses that domain-general signals in the ventral striatum and vmPFC (including subgenual ACC) were modulated by confidence level. These results are compatible with previous findings finding activity in the ventral striatum to be positively correlated with confidence (Daniel and Pollmann, 2012; Hebart et al., 2016; Guggenmos et al., 2016). Evidence of vmPFC encoding of confidence signals has been reported in connection with decision making and value judgments (De Martino et al., 2013; Lebreton et al., 2015). Our experimental design, however, does not allow us to disentangle whether the signals found in these regions pertain uniquely to confidence or if they are entangled with implicit value and reward signals (e.g., the expected value of being correct). Future experiments are needed to explicitly decouple reward from confidence to resolve this issue.

In our experimental design, perception and memory blocks were interleaved across runs, which raises the question as to whether the domain-specific neural substrates that we found would persist if subjects had to switch between tasks more often. Due to intertask "leaks" in confidence (Rahnev et al., 2015), in which confidence in one task influences confidence ratings on the following task (or even in the following trial), there is a possibility that interleaving blocks of different tasks might favor the observation of more domain-general confidence-related patterns.

Our experimental design assumes that visual perception and memory are distinct domains. We acknowledge that distinguishing between cognitive domains or individuating perceptual modalities is not straightforward (Macpherson, 2011). For instance, different modalities (e.g., vision, audition, touch, etc.), different aspects within a single modality (e.g., motion and color within vision), or closely related modalities (e.g., visual perception vs visual short-term memory) could be part of a unified perceptual domain for metacognitive purposes. Recent findings suggest that metacognitive efficiency in one perceptual modality predicts metacognitive efficiency in others and that they share electrophysiological markers (Faivre et al., 2018). Metacognitive efficiency is also correlated across vision and visual short-term memory, especially for features such as orientation (Samaha and Postle, 2017), and dACC and insula regions similar to those identified here have been found to show univariate confidence signals across both color and motion tasks in the visual domain (Heereman et al., 2015). However, it is an open question whether more fine-grained, modality-specific patterns of metacognitive activity could be decoded using multivariate approaches. More research is needed on the neural architecture of metacognition in other

cognitive domains and whether this architecture changes in a graded or discrete fashion as a function of task or stimulus.

In summary, our results provide evidence for the coexistence of content-rich metacognitive representations in anterior PFC with generic confidence-related signals in frontoparietal and cingulo-opercular regions. Such an architecture may be appropriate for "tagging" lower-level feelings of confidence with higher-order contextual information to allow effective behavioral control. Previous studies have tended to draw conclusions about either domain-specific or domain-general aspects of metacognition. Here, we reconcile these perspectives by demonstrating that both domain-specific and domain-general signals coexist in the human brain, thus laying the groundwork for a mechanistic understanding of reflective judgments of cognition.

References

Ais J, Zylberberg A, Barttfeld P, Sigman M (2016) Individual consistency in the accuracy and distribution of confidence judgments. Cognition 146: 377–386. CrossRef Medline

Andersson JL, Hutton C, Ashburner J, Turner R, Friston K (2001) Modeling geometric deformations in EPI time series. Neuroimage 13:903–919. CrossRef Medline

Ashburner J, Friston KJ (2005) Unified segmentation. Neuroimage 26:839–851. CrossRef Medline

Baird B, Smallwood J, Gorgolewski KJ, Margulies DS (2013) Medial and lateral networks in anterior prefrontal cortex support metacognitive ability for memory and perception. J Neurosci 33:16657–16665. CrossRef Medline

Bhandari A, Gagne C, Badre D (2017) Just above chance: is it harder to decode information from human prefrontal cortex BOLD signals? bioRxiv: 1–28. CrossRef

Boldt A, Yeung N (2015) Shared neural markers of decision confidence and error detection. J Neurosci 35:3478–3484. CrossRef Medline

Botvinick MM, Cohen JD, Carter CS (2004) Conflict monitoring and anterior cingulate cortex: an update. Trends Cogn Sci 8:539–546. CrossRef Medline

Brainard DH (1997) The Psychophysics Toolbox. Spatial Vision 10:433–436. CrossRef

Busey TA, Tunnicliff J, Loftus GR, Loftus EF (2000) Accounts of the confidence-accuracy relation in recognition memory. Psychon Bull Rev 7:26–48. Medline

Chua EF, Pergolizzi D, Weintraub RR (2014) The cognitive neuroscience of metamemory monitoring: understanding metamemory processes, subjective levels expressed, and metacognitive accuracy. In: The cognitive neuroscience of metacognition (Fleming SM, Frith CD, eds), pp 267–291. Berlin, Germany: Springer.

Cole MW, Reynolds JR, Power JD, Repovs G, Anticevic A, Braver TS (2013) Multi-task connectivity reveals flexible hubs for adaptive task control. Nat Neurosci 16:1348–1355. CrossRef Medline

Cole MW, Ito T, Braver TS (2016) The behavioral relevance of task information in human prefrontal cortex. Cereb Cortex 26:2497–2505. CrossRef Medline

Cortese A, Amano K, Koizumi A, Kawato M, Lau H (2016) Multivoxel neurofeedback selectively modulates confidence without changing perceptual performance. Nat Commun 7:13669. CrossRef Medline

Daniel R, Pollmann S (2012) Striatal activations signal prediction errors on confidence in the absence of external feedback. Neuroimage 59:3457–3467. CrossRef Medline

de Gardelle V, Mamassian P (2014) Does confidence use a common currency across two visual tasks? Psychol Sci 25:1286–1288. CrossRef Medline

De Martino B, Fleming SM, Garrett N, Dolan RJ (2013) Confidence in value-based choice. Nat Neurosci 16:105–110. CrossRef Medline

Donoso M, Collins AG, Koechlin E (2014) Foundations of human reasoning in the prefrontal cortex. Science 344:1481–1486. CrossRef Medline

Faivre N, Filevich E, Solovey G, Kühn S, Blanke O (2018) Behavioral, modeling, and electrophysiological evidence for supramodality in human metacognition. J Neurosci 38:263–277. Medline

Flavell JH (1979) Metacognition and cognitive monitoring: a new area of cognitive–developmental inquiry. Am Psychol 34:906–911. CrossRef

Fleck MS, Daselaar SM, Dobbins IG, Cabeza R (2006) Role of prefrontal and anterior cingulate regions in decision-making processes shared by memory and nonmemory tasks. Cereb Cortex 16:1623–1630. CrossRef Medline

Fleming SM (2017) Hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. Neuroscience of Consciousness 2017: 1–14. CrossRef

Fleming SM, Daw ND (2017) Self-evaluation of decision-making: a general bayesian framework for metacognitive computation. Psychol Rev 124:91–114. CrossRef Medline

Fleming SM, Lau HC (2014) How to measure metacognition. Front Hum Neurosci 8:443. CrossRef Medline

Fleming SM, Weil RS, Nagy Z, Dolan RJ, Rees G (2010) Relating introspective accuracy to individual differences in brain structure. Science 329: 1541–1543. CrossRef Medline

Fleming SM, Dolan RJ, Frith CD (2012a) Metacognition: computation, biology and function. Philos Trans R Soc Lond B Biol Sci 367:1280–1286. CrossRef Medline

Fleming SM, Huijgen J, Dolan RJ (2012b) Prefrontal contributions to metacognition in perceptual decision making. J Neurosci 32:6117–6125. CrossRef Medline

Fleming SM, Ryu J, Golfinos JG, Blackmon KE (2014) Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. Brain 137:2811–2822. CrossRef Medline

Galvin SJ, Podd JV, Drga V, Whitmore J (2003) Type 2 tasks in the theory of signal detectability: discrimination between correct and incorrect decisions. Psychon Bull Rev 10:843–876. CrossRef Medline

García-Pérez MA (1998) Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties. Vision Res 38:1861–1881. CrossRef Medline

Gehring WJ, Goss B, Coles MGH, Meyer DE, Donchin E (1993) A neural system for error detection and compensation. Psychol Sci 4:385–390. CrossRef

Gorgolewski KJ, Varoquaux G, Rivera G, Schwarz Y, Ghosh SS, Maumet C, Sochat VV, Nichols TE, Poldrack RA, Poline JB, Yarkoni T, Margulies DS (2015) NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. Front Neuroinform 9:8. CrossRef Medline

Guggenmos M, Wilbertz G, Hebart MN, Sterzer P (2016) Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. eLife 5:e13388. CrossRef Medline

Hebart MN, Görgen K, Haynes JD (2014) The decoding toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. Front Neuroinform 8:88. CrossRef Medline

Hebart MN, Schriever Y, Donner TH, Haynes JD (2016) The relationship between perceptual decision variables and confidence in the human brain. Cereb Cortex 26:118–130. CrossRef Medline

Heekeren HR, Marrett S, Ruff DA, Bandettini PA, Ungerleider LG (2006) Involvement of human left dorsolateral prefrontal cortex in perceptual decision making is independent of response modality. Proc Natl Acad Sci U S A 103:10023–10028. CrossRef Medline

Heereman J, Walter H, Heekeren HR (2015) A task-independent neural representation of subjective certainty in visual perception. Front Hum Neurosci 9:551. CrossRef Medline

Hilgenstock R, Weiss T, Witte OW (2014) You'd better think twice: post-decision perceptual confidence. Neuroimage 99:323–331. CrossRef Medline

Holroyd CB, Yeung N, Coles MG, Cohen JD (2005) A mechanism for error detection in speeded response time tasks. J Exp Psychol Gen 134:163–191. CrossRef Medline

Insabato A, Pannunzi M, Rolls ET, Deco G (2010) Confidence-related decision making. J Neurophysiol 104:539–547. CrossRef Medline

Kamitani Y, Sawahata Y (2010) Spatial smoothing hurts localization but not information: pitfalls for brain mappers. Neuroimage 49:1949–1952. CrossRef Medline

Kelemen WL, Frost PJ, Weaver CA 3rd (2000) Individual differences in metacognition: evidence against a general metacognitive ability. Mem Cognit 28:92–107. CrossRef Medline

Kerns JG, Cohen JD, Stenger VA, Carter CS (2004) Prefrontal cortex guides context-appropriate responding during language production. Neuron 43: 283–291. CrossRef Medline

Kiani R, Shadlen MN (2009) Representation of confidence associated with a decision by neurons in the parietal cortex. Science 324:759–764. CrossRef Medline

Kruschke J (2010) Doing Bayesian data analysis. San Diego, CA: Academic.

Lebreton M, Abitbol R, Daunizeau J, Pessiglione M (2015) Automatic integration of confidence in the brain valuation signal. Nat Neurosci 18: 1159–1167. CrossRef Medline

Macpherson F (2011) Introduction: individuating the senses. In: The senses (Macpherson F, ed), pp 3–42. Oxford, UK: OUP.

Maniscalco B, Lau H (2012) A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. Conscious Cogn 21:422–430. CrossRef Medline

Maniscalco B, McCurdy LY, Odegaard B, Lau H (2017) Limited cognitive resources explain a trade-off between perceptual and metacognitive vigilance. J Neurosci 37:1213–1224. CrossRef Medline

McCurdy LY, Maniscalco B, Metcalfe J, Liu KY, de Lange FP, Lau H (2013) Anatomical coupling between distinct metacognitive systems for memory and visual perception. J Neurosci 33:1897–1906. CrossRef Medline

Mendoza-Halliday D, Martinez-Trujillo JC (2017) Neuronal population coding of perceived and memorized visual features in the lateral prefrontal cortex. Nat Commun 8:15471. CrossRef Medline

Metcalfe J, Shimamura AP eds (1994) Metacognition. Cambridge, MA: MIT.

Nelson DL, McEvoy CL, Schreiber TA (2004) The University of South Florida free association, rhyme, and word fragment norms. Behav Res Methods Instrum Comput 36:402–407. CrossRef Medline

Nelson TO, Narens L (1990) Metamemory: a theoretical framework and new findings. In: The psychology of learning and motivation (Bower G, ed), pp 125–173. New York, NY: Academic.

Op de Beeck HP (2010) Against hyperacuity in brain reading: spatial smoothing does not hurt multivariate fMRI analyses? Neuroimage 49: 1943–1948. CrossRef Medline

Overgaard M, Sandberg K (2012) Kinds of access: different methods for report reveal different kinds of metacognitive access. Philos Trans R Soc Lond B Biol Sci 367:1287–1296. CrossRef Medline

Poldrack RA, Baker CI, Durnez J, Gorgolewski KJ, Matthews PM, Munafò MR, Nichols TE, Poline JB, Vul E, Yarkoni T (2017) Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat Rev Neurosci 18:115–126. CrossRef Medline

Pouget A, Drugowitsch J, Kepecs A (2016) Confidence and certainty: distinct probabilistic quantities for different goals. Nat Neurosci 19:366–374. CrossRef Medline

Purcell BA, Kiani R (2016) Hierarchical decision processes that operate over distinct timescales underlie choice and changes in strategy. Proc Natl Acad Sci U S A 113:E4531–E4540. CrossRef Medline

Rahnev D, Koizumi A, McCurdy LY, D'Esposito M, Lau H (2015) Confidence leak in perceptual decision making. Psychol Sci 26:1664–1680. CrossRef Medline

R Studio Team (2015) RStudio: Integrated Development for R. Available from http://www.rstudio.com.

Ridderinkhof KR, Ullsperger M, Crone EA, Nieuwenhuis S (2004) The role of the medial frontal cortex in cognitive control. Science 306:443–447. CrossRef Medline

Rounis E, Maniscalco B, Rothwell JC, Passingham RE, Lau H (2010) Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. Cogn Neurosci 1:165–175. CrossRef Medline

Ruby E, Giles N, Lau H (2017) Finding domain-general metacognitive mechanisms requires using appropriate tasks. bioRxiv: 1–20. CrossRef

Samaha J, Postle BR (2017) Correlated individual differences suggest a common mechanism underlying metacognition in visual perception and visual short-term memory. Proc Biol Sci 284:20172035. CrossRef Medline

Scheffers MK, Coles MG (2000) Performance monitoring in a confusing world: error-related brain activity, judgments of response accuracy, and types of errors. J Exp Psychol Hum Percept Perform 26:141–151. CrossRef Medline

Schnyer DM, Verfaellie M, Alexander MP, LaFleche G, Nicholls L, Kaszniak AW (2004) A role for right medial prefrontal cortex in accurate feeling-of-knowing judgments: evidence from patients with lesions to frontal cortex. Neuropsychologia 42:957–966. CrossRef Medline

Shimamura AP, Squire LR (1986) Memory and metamemory: a study of the feeling-of-knowing phenomenon in amnesic patients. J Exp Psychol Learn Mem Cogn 12:452–460. CrossRef Medline

Simons JS, Peers PV, Mazuz YS, Berryhill ME, Olson IR (2010) Dissociation between memory accuracy and memory confidence following bilateral parietal lesions. Cereb Cortex 20:479–485. CrossRef Medline

Sladky R, Friston KJ, Tröstl J, Cunnington R, Moser E, Windischberger C (2011) Slice-timing effects and their correction in functional MRI. Neuroimage 58:588–594. CrossRef Medline

Tang H, Yu HY, Chou CC, Crone NE, Madsen JR, Anderson WS, Kreiman G (2016) Cascade of neural processing orchestrates cognitive control in human frontal cortex. eLife 5:e12352. CrossRef Medline

Vo VA, Li R, Kornell N, Pouget A, Cantlon JF (2014) Young children bet on their numerical skills: metacognition in the numerical domain. Psychol Sci 25:1712–1721. CrossRef Medline

Wilson M (1988) MRC psycholinguistic database: machine-usable dictionary, version 2.00. Behav Res Methods Instrum Comput 20:6–10. CrossRef

Yokoyama O, Miura N, Watanabe J, Takemoto A, Uchida S, Sugiura M, Horie K, Sato S, Kawashima R, Nakamura K (2010) Right frontopolar cortex activity correlates with reliability of retrospective rating of confidence in short-term recognition memory performance. Neurosci Res 68:199–206. CrossRef Medline